- Ill conditioned Problem (multicollinearity (X^TX) is singular) : $\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$
- To fix it we use Ridge, LASSO or ElasticNet

* **Ridge Regression Method** :

L^2 loss = $\sum_{i=1}^{n} e_i^2$ = SSE

L^1 loss = $\sum_{i=1}^{n} |e_i|$ Not differentiable

$\hat{\beta}_{OLS} = \min_{\beta}(L^2 \text{ loss})$ * so we get $\hat{\beta}_{OLS}$

Penalty terms ensures that $(X^TX)^{-1}$ always exists

* is reformulated to take loss + penalty approach that attends to shrink the coefficients (some of the parameters are close to zero) by imposing a penalty on their size (size of the coefficients).

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$

$$L^2 \text{ loss} = \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j)^2$$

Two formulations were proposed :

$$\hat{\beta}_{Ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

loss → penalty

$$\hat{\beta}_{Ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} ; \sum_{j=1}^{p} \beta_j^2 \leq t , \; t \neq \lambda$$

λ := shrinkage parameter estimated by cross validation.

$$\lambda = 0 \Rightarrow \hat{\beta}_{Ridge} = \hat{\beta}_{OLS}$$

Read about Time varying Ridge Model

* Estimation of $\hat{\beta}$:

$$\min ((Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta) =$$
$$\min ((Y^T - \beta^T X^T)(Y - X\beta) + \lambda \beta^T \beta) =$$
$$\min (Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \lambda \beta^T \beta) =$$
$$\min (Y^T Y - \beta^T X^T Y - (\beta^T X^T Y)^T + \beta^T X^T X\beta + \lambda \beta^T \beta) =$$
$$\min (Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta + \lambda \beta^T \beta)$$

$$\frac{\partial}{\partial \beta} = -2X^T Y + 2X^T X\beta + 2\lambda\beta = 0$$

$$X^T X\beta + \lambda\beta - X^T Y = 0$$
$$\beta (X^T X + \lambda I) = X^T Y$$

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

* The coefficients can be estimated by OLS & λ is estimated by cross validation.

* The choice of quadratic penalty adds a +ve constant to the diagonal element of (X^TX) and that mitigates the problem of multicollinearity.

* $\hat{\beta}_{Ridge}$ is linear

* $$\begin{aligned} E(\hat{\beta}_{Ridge}) &= E((X^TX + \lambda I)^{-1} X^T Y) \\ &= (X^TX + \lambda I)^{-1} X^T E(X\beta + \varepsilon) \\ &= (X^TX + \lambda I)^{-1} X^T X \beta \\ &\neq \beta \end{aligned}$$

Hence, it is a baised estimater but in the cost of bias, ridge reduces the variance

$$\begin{aligned} Var(\hat{\beta}_{Ridge}) &= Var((X^TX + \lambda I)^{-1} X^T Y) \\ &= ((X^TX + \lambda I)^{-1} X^T)((X^TX + \lambda I)^{-1} X^T)^T Var(Y) \\ &= \sigma^2 (X^TX + \lambda I)^{-1} X^T X (X^TX + \lambda I)^{-1} \end{aligned}$$

* **LASSO Regression** :

LASSO : least absolute shrinkage & selection operator

In Ridge the penalty is $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks all $\beta_j \to 0$ but LASSO solves this problem.

LASSO uses L_1 norm that has the effect of forcing some coefficients to become zero which leads to sparse model.

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2\right);\ \sum_{j=1}^{p}|\beta_j|\le t,\ t\to\lambda$$

λ is estimated by cross validation.

$\hat{\beta}_{LASSO}$ does not have a closed solution like OLS & Ridge but it is helpful when we have 1000 features but only 200 samples as an example.

We can use package to do this year!

Also, check Bayesian LASSO, Polynomial, Non linear and L_0 regularization

* **Elastic Net Method** :

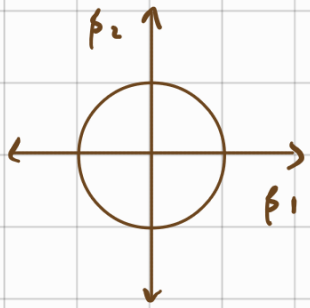
$$\hat{\beta}_{EN}=\text{Loss}+\lambda\sum_{j=1}^{p}(\alpha\beta_j^2+(1-\alpha)|\beta_j|)$$

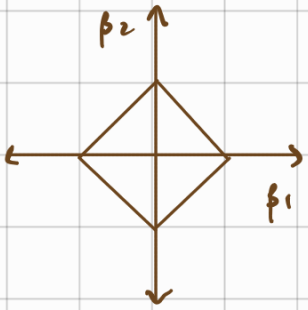
so it is a compromise between Ridge & LASSO

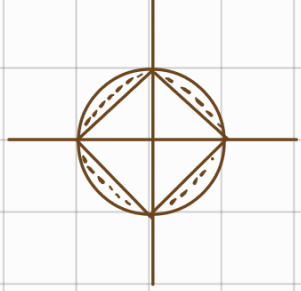
Ridge LASSO Elastic Net

$\beta_1^2 + \beta_2^2 \leq 1$ $\quad |\beta_1| + |\beta_2| \leq 1$

$u^2 + y^2 \leq 1$ $\quad |u| + |y| \leq 1$







* Remark: Before fitting regularization models (LASSO, Ridge) the predictors X's should be standardized (unless using a package)

$$X_{new} = \frac{X - mean}{sd}$$

because LSE is a scaling variant but this is not.

glmnet standardize by default

* Standard error: $\sqrt{MS_{Res}}$

* We do cross validation to estimate alpha and lambda for elastic net in R for best results (10 fold CV).

* Transformations:

	X	Y	$\hat{Y}$
Train	$\log(X_{Tr})$	$\log(Y_{Tr})$	
Test	$\log(X_{Test})$	X	$\hat{y} \longrightarrow e^{\hat{y}}$

Reverse transform

* Standardization Comparison : $\frac{x - \mu_x}{\sigma_x}$

Train : $\frac{x_{Tr} - \mu_{x_{Tr}}}{\sigma_{x_{Tr}}}$ Test : $\frac{x_{Ts} - \mu_{Tr}}{\sigma_{Tr}}$

don't use μ_{Ts} / σ_{Ts} otherwise we get data leakage

High Range / Not following assumptions :

$$y' \longrightarrow \frac{y - \mu_y}{\sigma_y}$$

$$y'_{Tr} = \frac{y_{Tr} - \mu_{yTr}}{\sigma_{yTr}}$$

$$y'_{Ts} = \frac{y_{Ts} - \mu_{yTr}}{\sigma_{yTs}}$$